


Endpoint operations

RunPod's endpoints facilitate submitting jobs and retrieving outputs.

To use these endpoints, you will need to have your endpoint ID. The constructed URL will start with `https://api.runpod.ai/v2/{endpoint_id}/{operation}` followed by an operation.

Operations available to all users are:

- `/run`: Asynchronous endpoint for submitting jobs. Returns a unique Job ID.
 - Payload capacity: 10 MB
 - Rate limit: 1000 per second
 - Job availability: Successful job results are accessible for 30 minutes after completion
- `/runsync`: Synchronous endpoint for shorter running jobs, returning immediate results.
 - Payload capacity: 20 MB
 - Rate limit: 2000 per second
 - Job availability: Successful job results are accessible for 60 seconds after completion
- `/stream/{job_id}`: For streaming results from generator-type handlers.
- `/status/{job_id}`: To check the job status and retrieve outputs upon completion.
- `/cancel/{job_id}`: To cancel a job prematurely.
- `/health`: Provides worker statistics and endpoint health.
 - Only accepts `GET` methods
- `/purge-queue`: Clears all queued jobs, it will not cancel jobs in progress.

 [Edit this page](#)

Previous

« [Endpoint configurations](#)

Next

[Job states](#) »



Ask AI


[Overview](#)

[Tutorials](#)

[AI APIs](#)


Community

[Discord](#) 

[Contact us](#) 

More

[Blog](#) 

[GitHub](#) 

Copyright © 2024 RunPod



Ask AI